Supplementary Material to Learning with Known Operators reduces Maximum Training Error Bounds

Andreas K. Maier, Christopher Syben, Bernhard Stimpel,
Tobias Würfl, Mathis Hoffmann, Frank Schebesch, Weilin Fu,
Leonid Mill[*], Lasse Kling[†], and Silke Christiansen[‡]

July 3, 2019

1 Details on Experiments on Deep Learning Computed Tomography

Our method was evaluated on the publicly available dataset from the Low Dose CT Grand Challenge [1]. We simulated incomplete projection data for 2378 slices of 10 different patients in parallel-beam and fan-beam geometry. The slices had a resolution of 512×512 pixels. A training sample for our experiment in parallel-beam geometry consisted of a groundtruth slice as label and projection data from an angular range of 175 degree with an angular increment of one degree as input to the neural network. For the fan-beam experiment we used projection data from an angular range of 180 degree also with an increment of one degree. Separation of training and test data was performed in a ten-fold cross validation. As evaluation measure we computed the root-mean-square error divided by the maximum value (rRMSE) of the slice averaged over all ten folds.

In our parallel-beam experiment, we set the reconstruction filter as a 2D filter combining information over 5 slices but we initialize only its central elements to the Ramachandran-Lakshminarayan filter discretization. All other weights are initialized to zero to start learning with the classical analytic algorithm. After the reconstruction we use a maxout [2] unit, performing an element-wise maximum operation over four 3×3 filters which are finally combined by a

[*]A.K. Maier, C. Syben, T Würfl, M. Hoffmann, F. Schebesch, W. Fu, B. Stimpel, and L. Mill are with the Department of Computer Science, Friedrich-Alexander University Erlangen-Nürnberg, Germany. E-mail: andreas.maier@fau.de.

[†]L. Kling is with the Helmholtz Zentrum Berlin für Materialien und Energie, Germany.

[‡]S.Christiansen is with the Physics Department, Free University Berlin and the Helmholtz Zentrum Berlin für Materialien und Energie, Germany.

convolution of a single 3×3 filter. All those convolutional layers are initialized from a Gaussian distribution using a mean and standard deviation of 10^{-3}. We used the "RMSPROP" solver with a batch size of one, a learning rate of 10^{-6} and its decay parameter β set to $2 \cdot 10^{-2}$. The reconstruction filter received a small learning rate of 10^{-15}.

We trained the network for a fixed number of 10000 iterations while decreasing the learning rate by 4% every 100 iterations and compared our method to the results of the filtered backprojection algorithm. Our trained algorithm achieved an error of $3.54 \cdot 10^{-3}\%$, outperforming the analytic algorithm with a rRMSE of $6.78 \cdot 10^{-3}\%$.

In our fan-beam experiment, we used the introduced architecture which is equivalent to filtered backprojection in fan-beam geometry and only trained the weighting layer $\mathbf{W}$ using 10000 iterations of the "ADAGRAD" solver, a learning rate of $2 \cdot 10^{-2}$ and a batch size of one. $\mathbf{K}$ was not altered and initialized with the filter weights from the discretized analytical solution. The weighting layer was initialized with appropriate cosine weights and the conventional Parker weights for short scans [3]. Despite these few trainable parameters our trained method reduced the rRMSE of $5.31 \cdot 10^{-3}\%$ using the analytic method to $3.92 \cdot 10^{-3}\%$
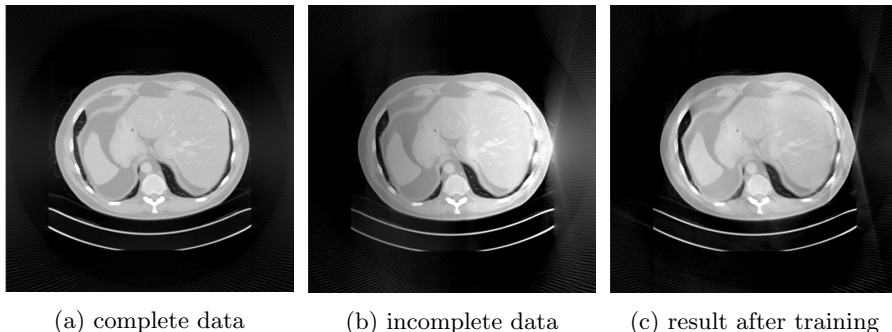

(a) complete data	(b) incomplete data	(c) result after training

Figure 1: Comparison of different reconstruction methods in fan-beam geometry: (a) shows a reconstruction from a complete data set collected over $200°$ of rotational range. (b) is reconstructed from $180°$ of data and already exhibits strong artifacts while (c) learns to compensate for the missing data.

2   Details on Experiments on Learning from Heuristic Algorithms

Our algorithms are evaluated on the Digital Retinal Images for Vessel Extraction (DRIVE) [4] database. The database contains 40 8-bit RGB photographs of size 565×584, and is evenly divided into one training and one testing set. We further randomly select four datasets from the training set for the validation during the training process. All images are preprocessed using a pipeline including

green channel extraction, CLAHE [5] and normalization. For a fair performance evaluation, provided field of view masks are eroded inward with four pixels to completely eliminate the border effect.

The Frangi filter is implemented over 8 Gaussian scales with σ ranging from 0.5 to 4.0. The blobness factor β and structureness factor c of all scales are set to 0.5 and 1.0, respectively. The Frangi-Net is initialized exactly the same as the Frangi filter. To enable end-to-end training, the multi-scale vesselness map is followed by two convolution layers with 1×1 filters to normalize the data range and generate the double-channel softmax output. Overall, Frangi-Net contains 6526 trainable parameters.

A trainable guided filter block [6] is further employed to preprocess images before the Frangi-Net. The additional block employs a Context Aggregation Network (CAN) [7] with five levels and eight features channels to generate the guidance map, and a small CNN with two convolutional layers to extract features from the input images. Overall, the guided filter block introduces 3050 more parameters into the Frangi-Net workflow.

The U-Net adopted in this work has three levels and 16 channels in the initial filter. On the basis of the original U-Net configuration, a batch normalization layer is added after each convolution layer to stabilize the training process, and the deconvolution layers in the expanding path are replaced with upsampling layers followed by a 1×1 convolution layer to avoid the checkerboard effect [8]. In total, the U-Net contains 111536 trainable parameters.

The objective function for both networks is constructed by multiplying a weighting map which emphasizes thin vessels to the class-balanced focal loss [9] with modulating factor of 2.0. Additionally for the U-Net, a L2-norm regularizer of scale 0.05 is employed to alleviate the overfitting problem. The weighting map w is created with manual labels using $w = \max(\frac{1}{0.18 \times d}, 1)$, where d denotes the tube diameter in the manual label. Adam optimizer is selected for objective function minimization. The learning rates are initialized to 10^{-4} for Frangi-Net pipelines and 5×10^{-4} for the U-Net, and decay after each 2,000 iterations by 10%. For Frangi-Net pipelines, each training batch is composed of 200 image patches of size 88×88. For U-Net, each batch contains 50 patches of shape 168×168. Data augmentation methods including rotation, shearing, Gaussian noise and intensity shifting are employed for better generalization of the networks.

The performance comparison of Frangi filter, two Frangi-Net workflows and the U-Net is presented in Table. 1. Output vesselness or probability maps are binarized with a single threshold which maximizes the F1 score of the validation set. The Frangi-Net outperforms the original Frangi filter for all evaluation metrics, achieving an AUC score around 0.96 with fewer than 6% of the U-Net parameters. With the trainable guided filter block for preprocessing, Frangi-Net achieves on-par performance as the U-Net with fewer than 9% of the U-Net unknown weights.

Table 1: Performance evaluations of Frangi filter (FF), Frangi-Net (FN), Guided filter layer + Frangi-Net (GF) and U-Net (UN) on the DRIVE testing sets.

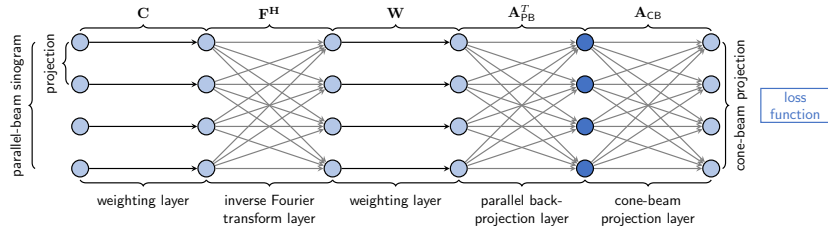|	specificity	sensitivity	F1 score	accuracy	AUC
FF	.9616±.0150	.7528±.0612	.7477±.0323	.9341±.0089	.9401
FN	.9633±.0125	.8008±.0590	.7812±.0256	.9419±.0070	.9610
GF	.9729±.0060	.7982±.0546	.8048±.0191	.9498±.0048	.9719
UN	.9756±.0057	.7942±.0576	.8097±.0227	.9516±.0056	.9743

Figure 2: The derived rebinning network architecture. As $\mathbf{C}$ is a diagonal matrix in Fourier domain it describes a convolution followed by a diagonal matrix $\mathbf{W}$ in spatial domain, which is merely a point-wise multiplication. The filtered and weighted projections are fed into a parallel-beam back-projection $\mathbf{A}_{\mathrm{PB}}^{\top}$ with a subsequent cone-beam forward-projection $\mathbf{A}_{\mathrm{CB}}$ to obtain the projection under the target geometry. Light blue nodes represent intermediate results in projection domain, while dark blue nodes stands for volume domain.

3 Details on Experiments on Deriving Networks

Analog to the case of CT reconstruction, we can now map above formula into a deep network and train the remaining parameters using back-propagation as displayed in Figure 2.

The approximations for $\mathbf{C}$ permit two assumptions: $\mathbf{C}$ independent of the projection angle or a projection dependent filter, i.e. one filter per parallel projection. In the following we consider the results of the projection dependent filter, for detailed discussion about both assumptions we refer to [10]. The layers $\mathbf{A}$ and $\mathbf{A}^{\top}$ are as in the first experiment directly implemented algorithms and can be treated as fully connected layers with fixed weights. The training of the network is conducted with 65 numerical phantoms with 256×256 pixel size for the parallel- to fan-beam rebinning geometry. The phantoms, displayed in Figure 3, are designed to bring certain properties in the training process. We want to highlight the properties of phantoms in Figure 3a and 3b, which are designed to be sensitive to perspective distortion, causing strong gradients in the training process. To ensure gradients for the full Fourier spectrum 50 noise phantoms are included into the training dataset. The training dataset includes also circles with varying radii. To evaluate the learned algorithm on
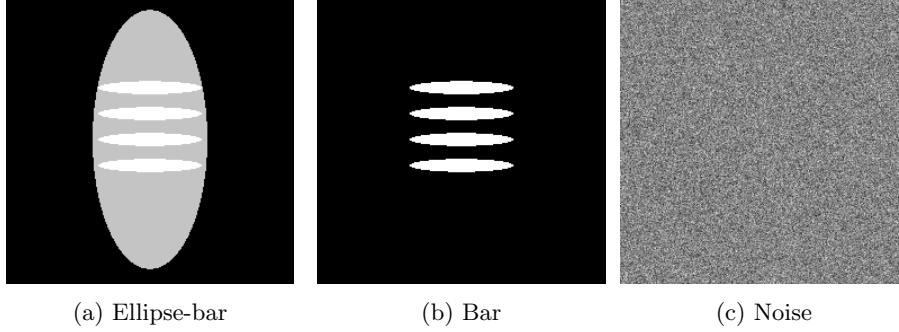
(a) Ellipse-bar   (b) Bar   (c) Noise

Figure 3: Numerical phantom types used in training. Circles with varying radii and Ellipses are used to bring mass and homogeneous areas into the training. The properties of the Ellipse-bar (a) and Bars (b) phantoms, both with varying number of bars causing strong errors for the perspective distortion and the last type is normal distributed noise (c) to cover the entire Fourier spectrum.

the Shepp-Logan phantom [11] the mean squared error (MSE) is computed and additionally we demonstrated the trained algorithm on real parallel-beam MR projections and computed the distorted stacked fan-beam projection. The fan-beam projection of the Shepp-Logan phantom of the learned algorithm shows smaller errors especially at the edges and high frequency parts, which is reflected in the lower MSE value compared to the analytical approach. For a detailed discussion about the result and the filter interpretations we refer to [10].

4   Proof of Known Output Operator Theorem

For the general case of $g(\mathbf{x})$ being an arbitrary Lipschitz-continuous function on compact set $\mathcal{S}$, we start with

$$
\begin{aligned}
f(\mathbf{x}) &= g(\mathbf{u}(\mathbf{x})) \\
&= g\left(\hat{\mathbf{u}}(\mathbf{x}) + \mathbf{e}_u\right).
\end{aligned}
$$

With the property of $g(\mathbf{x})$ being Lipschitz bound by l_g according to Theorem 1, we can relax its structure.

Theorem 1 (Lipschitz Bounds on L_p Norms). *Let $g(x) : \mathbb{R}^{N_D} \to \mathbb{R}$ be a continuous function. Then, there exists a constant $l_g \geq 0$ for which the inequalities*

$$
g(\mathbf{x} + \mathbf{e}) \leq g(\mathbf{x}) + l_g ||\mathbf{e}||_p \tag{1}
$$

$$
g(\mathbf{x} + \mathbf{e}) \geq g(\mathbf{x}) - l_g ||\mathbf{e}||_p. \tag{2}
$$

hold for $p \in \{1, 2\}$.

(a) Ground-truth
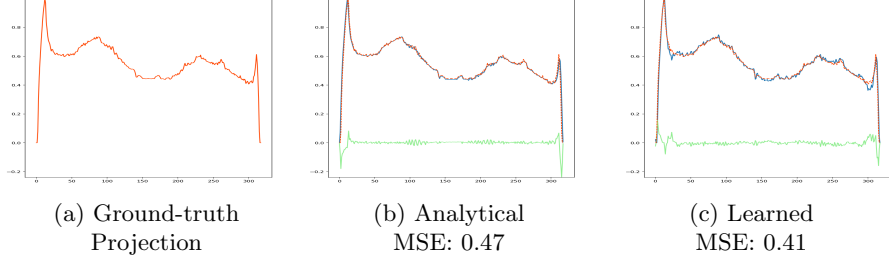Projection

(b) Analytical
MSE: 0.47

(c) Learned
MSE: 0.41

Figure 4: The 2D Shepp-Logan Phantom is used as a quantitative measurement, 7 parallel-beam projections according to [10] are generated and rebinned to one fan-beam projection. The ground-truth fan-beam projection is shown in (a). The analytical rebinning result is shown in (b) as the blue line, green is the difference to the ground-truth. The result of the newly derived algorithm is plotted in (c) as the blue line.

Now, we use Theorem 1, Eq. 1:

$$
\begin{aligned}
g(\mathbf{x} + \mathbf{e}_u) & \leq g(\mathbf{x}) + l_g \cdot ||\mathbf{e}_u||_p \\
\underbrace{g(\mathbf{x} + \mathbf{e}_u) - g(\mathbf{x})} & \leq l_g \cdot ||\mathbf{e}_u||_p \\
e_u & \leq l_g \cdot ||\mathbf{e}_u||_p
\end{aligned}
$$

Theorem 1, Eq. 2 allows to find an analog lower bound which allows to compute

$$
|e_u| \leq \quad l_g \cdot ||\mathbf{e}_u||_p. \tag{3}
$$

$\square$

5 Proof of the Unknown Operator Theorem

We start with the definition of $f(\mathbf{x})$ and introduce the approximations of $g(\mathbf{x})$ and $\mathbf{u}(\mathbf{x})$, compensating for the errors e_g and $\mathbf{e}_u$:

$$
\begin{aligned}
f(\mathbf{x}) & = g(\mathbf{u}(\mathbf{x})) = \hat{g}(\mathbf{u}(\mathbf{x})) + e_g \\
& = \sum_{j=0}^{N_{\hat{g}}-1} g_j \varphi \left(\mathbf{w}_j^\top \mathbf{u}(\mathbf{x}) \right) + e_g \\
& = \sum_{j=0}^{N_{\hat{g}}-1} g_j \varphi \left(\mathbf{w}_j^\top (\hat{\mathbf{u}}(\mathbf{x}) + \mathbf{e}_u) \right) + e_g \\
& = \sum_{j=0}^{N_{\hat{g}}-1} g_j \varphi \left(\mathbf{w}_j^\top \hat{\mathbf{u}}(\mathbf{x}) + \underbrace{\mathbf{w}_j^\top \mathbf{e}_u}_{e_{\mathbf{u}_j}} \right) + e_g. \tag{4}
\end{aligned}
$$

6

In this configuration, we use the property of $\varphi(x)$ being Lipschitz-bounded by l_φ. However, as the activation function appears in a linear combination, we have to make use of a property that is derived form the Lipschitz bound.

Theorem 2 (Modified Lipschitz Bounds). *Let $\varphi(x)$ be a Lipschitz-bounded function with Lipschitz bound $l \geq 0$ and $e \in \mathbb{R}$ an arbitrary real number, then the following inequalities hold for all $g_j \in \mathbb{R}$.*

$$g_j \varphi(x + e) \leq g_j \varphi(x) + |g_j| \cdot l \cdot |e| \tag{5}$$
$$g_j \varphi(x + e) \geq g_j \varphi(x) - |g_j| \cdot l \cdot |e|. \tag{6}$$

Using Theorem 2 Eq. 5, we now arrive at the following inequality:

$$f(\mathbf{x}) \leq \underbrace{\sum_{j=0}^{N_{\hat{g}}-1} g_j \varphi(\mathbf{w}_j^\top \hat{\mathbf{u}}_j(\mathbf{x}))}_{=\hat{g}(\hat{\mathbf{u}}(\mathbf{x}))=\hat{f}(\mathbf{x})} + \sum_{j=0}^{N_{\hat{g}}-1} |g_j| \cdot l_\varphi \cdot |e_{\mathbf{u}_j}| + e_g$$

$$\leq \hat{f}(\mathbf{x}) + \sum_{j=0}^{N_{\hat{g}}-1} |g_j| \cdot l_\varphi \cdot |e_{\mathbf{u}_j}| + e_g.$$

Subtraction of $\hat{f}(\mathbf{x})$ yields

$$\underbrace{f(\mathbf{x}) - \hat{f}(\mathbf{x})} \leq \sum_{j=0}^{N_{\hat{g}}-1} |g_j| \cdot l_\varphi \cdot |e_{\mathbf{u}_j}| + e_g$$

$$e_f \leq \sum_{j=0}^{N_{\hat{g}}-1} |g_j| \cdot l_\varphi \cdot |e_{\mathbf{u}_j}| + e_g.$$

Now, using $|e_g| \leq \epsilon_g$ leads to

$$e_f \leq \sum_{j=0}^{N_{\hat{g}}-1} |g_j| \cdot l_\varphi \cdot |e_{\mathbf{u}_j}| + \epsilon_g.$$

Following the idea of the upper bound, an analog lower bound (cf. Theorem 2 Eq. 6) is found:

$$e_f \geq - \sum_{j=0}^{N_{\hat{g}}-1} |g_j| \cdot l_\varphi \cdot |e_{\mathbf{u}_j}| - \epsilon_g$$

Thus, an absolute bound for e_f is found as

$$|e_f| \leq \sum_{j=0}^{N_{\hat{g}}-1} |g_j| \cdot l_\varphi \cdot |e_{\mathbf{u}_j}| + \epsilon_g.$$

$\square$

Note that the results of the Known Output Operator Theorem and the Unknown Operator Theorem are not in contradiction. In fact there are also cases in which equality can be constructed. Obviously $g(\mathbf{x})$ has to be known for these cases, i.e. $e_g = 0$:

$$\sum_{j=0}^{N_{\hat{g}}-1} |g_j| \cdot l_\varphi \cdot |e_{\mathbf{u}_j}| = l_\varphi \quad \sum_{j=0}^{N_{\hat{g}}-1} |g_j| \cdot |e_{\mathbf{u}_j}|$$

$$= l_\varphi \quad \underbrace{\sum_{j=0}^{N_{\hat{g}}-1} |g_j \cdot \mathbf{w}_j^\top \mathbf{e}_u|}_{||\mathbf{G}\mathbf{e}_u||_1}$$

$$= l_\varphi \quad ||\mathbf{G}\mathbf{e}_u||_1 \qquad (7)$$

Here, we introduce a matrix $\mathbf{G}$ containing the vectors $\mathbf{w}_j$ scaled with g_j in each row. Thus for the case of a trivial g in the form of a general function approximator with $\mathbf{G} = \mathbf{I}$, $l_\varphi = l_g$, and $p = 1$, equality for both theorems is constructed. A more interesting relation is constructed using the an upper bound for Eq. 7

$$l_\varphi \cdot ||\mathbf{G}\mathbf{e}_u||_1 \le l_\varphi \cdot ||\mathbf{G}|| \cdot ||\mathbf{e}_u||_1$$

for a general matrix norm $||\mathbf{G}||$ consistent with $|| \cdot ||_1$. Now, the Known Output Operator Theorem becomes an upper bound for the Unknown Operator Theorem for functions g in the form of universal function approximators with $||\mathbf{G}|| \cdot l_\varphi = l_g$, and $p = 1$, as the following equality becomes true

$$l_\varphi \cdot ||\mathbf{G}|| \cdot ||\mathbf{e}_u||_1 = l_g \cdot ||\mathbf{e}_u||_1. \qquad (8)$$

6 Proof for Unknown Operators in Deep Networks

The proof for this bound is found by mathematical induction. We start with the upper bound on $\mathbf{e}_{\mathbf{f},\ell}$ for component k:

$$e_{f,\ell,k} \le \sum_{\ell_i=1}^{\ell} ||\mathbf{e}_{u,\ell_i}||_p \cdot l_{\mathbf{f}_{\ell_i-1}} \qquad (9)$$

Using the definition of $e_{f,\ell,k} = [\mathbf{f}_\ell(\mathbf{x}_\ell)]_k - [\hat{\mathbf{f}}_\ell(\mathbf{x}_\ell)]_k$, we get

$$[\mathbf{f}_\ell(\mathbf{x}_\ell)]_k \le [\hat{\mathbf{f}}_\ell(\mathbf{x}_\ell)]_k + \sum_{\ell_i=1}^{\ell} ||\mathbf{e}_{u,\ell_i}||_p \cdot l_{\mathbf{f}_{\ell_i-1}} \qquad (10)$$

The base case $\ell = 1$ results in the multi-dimensional expansion of the Universal Approximation Theorem shown in the main paper and is therewith true.

Therefore, we need to demonstrate the inductive step setting $\ell := \ell + 1$. With $\mathbf{f}_{\ell+1}(\mathbf{x}_{\ell+1}) = \mathbf{f}_\ell(\mathbf{u}_{\ell+1}(\mathbf{x}_{\ell+1}))$ an additional layer is introduced into the network. Approximation of $\mathbf{u}_{\ell+1}$ now yields

$$\mathbf{f}_\ell(\mathbf{u}_{\ell+1}(\mathbf{x}_{\ell+1})) = \mathbf{f}_\ell(\hat{\mathbf{u}}_{\ell+1}(\mathbf{x}_{\ell+1}) + \mathbf{e}_{u,\ell+1}) \tag{11}$$

At this point, we can use Theorem 1 times to extract the error vector $\mathbf{e}_{u,\ell+1}$.

$$\mathbf{f}_\ell(\mathbf{u}_{\ell+1}(\mathbf{x}_{\ell+1})) \leq \mathbf{f}_\ell(\hat{\mathbf{u}}_{\ell+1}(\mathbf{x}_{\ell+1})) + ||\mathbf{e}_{u,\ell+1}||_p \cdot l_{\mathbf{f}_\ell}. \tag{12}$$

If we now use Eq. 10 to approximate $\mathbf{f}_\ell$, we arrive at

$$
\begin{aligned}
[\mathbf{f}_{\ell+1}(\mathbf{x}_{\ell+1})]_k &\leq [\hat{\mathbf{f}}_\ell(\hat{\mathbf{u}}_{\ell+1}(\mathbf{x}_{\ell+1}))]_k + ||\mathbf{e}_{u,\ell+1}||_p \cdot l_{\mathbf{f}_\ell} + \sum_{\ell_i=1}^{\ell} ||\mathbf{e}_{u,\ell_i}||_p \cdot l_{\mathbf{f}_{\ell_i}-1} \\
&\leq [\hat{\mathbf{f}}_{\ell+1}(\mathbf{x}_{\ell+1})]_k + \sum_{\ell_i=1}^{\ell+1} ||\mathbf{e}_{u,\ell_i}||_p \cdot l_{\mathbf{f}_{\ell_i}-1}.
\end{aligned}
$$

Subtraction of $[\hat{\mathbf{f}}_{\ell+1}(\mathbf{x}_{\ell+1})]_k$ finally results in

$$e_{f,\ell+1,k} \leq \sum_{\ell_i=1}^{\ell+1} ||\mathbf{e}_{u,\ell_i}||_p \cdot l_{\mathbf{f}_{\ell_i}-1}. \tag{13}$$

The derivation for the lower bound follows the same line just using the negative sign. $\qquad\square$

Analysis of the recursion the reveals the connection between Lipschitz constants $l_{\mathbf{f}_\ell}$ and $l_{\mathbf{u}_\ell}$. The Lipschitz bound of $\mathbf{f}_\ell$ is found as

$$||\mathbf{f}_\ell(\mathbf{x}_\ell) - \mathbf{f}_\ell(\mathbf{x}_\ell + \mathbf{e})|| \leq l_{\mathbf{f}_\ell}||\mathbf{x}_\ell - (\mathbf{x}_\ell + \mathbf{e})||. \tag{14}$$

Expansion of the recursion yields

$$\mathbf{f}_\ell(\mathbf{x}_\ell) - \mathbf{f}_\ell(\mathbf{x}_\ell + \mathbf{e}) = \mathbf{u}_1(\ldots(\mathbf{u}_\ell(\mathbf{x}_\ell))) - \mathbf{u}_1(\ldots(\mathbf{u}_\ell(\mathbf{x}_\ell + \mathbf{e}))). \tag{15}$$

Which is bounded by

$$
\begin{aligned}
||\mathbf{f}_\ell(\mathbf{x}_\ell) - \mathbf{f}_\ell(\mathbf{x}_\ell + \mathbf{e})|| &\leq l_{\mathbf{u}_1}||\mathbf{u}_2(\ldots(\mathbf{u}_\ell(\mathbf{x}_\ell))) - \mathbf{u}_2(\ldots(\mathbf{u}_\ell(\mathbf{x}_\ell + \mathbf{e})))|| \\
&\leq l_{\mathbf{u}_1} \cdot l_{\mathbf{u}_2}||\mathbf{u}_3(\ldots(\mathbf{u}_\ell(\mathbf{x}_\ell))) - \mathbf{u}_3(\ldots(\mathbf{u}_\ell(\mathbf{x}_\ell + \mathbf{e})))||. \\
&\vdots \\
&\leq \prod_{\ell_i=1}^{\ell} l_{\mathbf{u}_{\ell_i}}||\mathbf{x}_\ell - (\mathbf{x}_\ell + \mathbf{e})||. \tag{16}
\end{aligned}
$$

$l_{\mathbf{f}_\ell} \leq \prod_{\ell_i=1}^{\ell} l_{\mathbf{u}_{\ell_i}}$ follows.

7 Proof of Theorem 1

We start with generalised Lipschitz definition after [12], where $p \in \{1, 2\}$

$$
\begin{aligned}
|g(\mathbf{y}) - g(\mathbf{x})| &\leq l_g ||\mathbf{y} - \mathbf{x}||_p \\
-l_g ||\mathbf{y} - \mathbf{x}||_p \leq \quad g(\mathbf{y}) - g(\mathbf{x}) &\leq l_g ||\mathbf{y} - \mathbf{x}||_p \\
-l_g ||\mathbf{y} - \mathbf{x}||_p + g(\mathbf{x}) \leq \qquad g(\mathbf{y}) \qquad &\leq l_g ||\mathbf{y} - \mathbf{x}||_p + g(\mathbf{x})
\end{aligned}
$$

By substituting $\mathbf{y} := \mathbf{x} + \mathbf{e}$, the right inequality yields Eq. 1 and the left inequality Eq. 2. $\qquad\square$

8 Proof of Theorem 2

The Lipschitz condition for continuous functions $\varphi(x)$ is given as

$$
|\varphi(x + e) - \varphi(x)| \leq l|e|
$$

which is equivalent to

$$
\begin{aligned}
-l|e| \leq \quad \varphi(x + e) - \varphi(x) &\leq l|e| \\
\varphi(x) - l|e| \leq \qquad \varphi(x + e) \qquad &\leq \varphi(x) + l|e|
\end{aligned}
$$

Next, we multiply with $g_j \in \mathbb{R} \setminus \{0\}$. Note that for $g_j = 0$, the inequality is trivial but not sufficient for the Lipschitz continuity. Now, we investigate the resulting two cases:

- $g > 0$: The inequality becomes

$$
g_j \varphi(x) - g_j l|e| \leq g_j \varphi(x + e) \leq g_j \varphi(x) + g_j l|e|.
$$

  and the right inequality can be rewritten as

$$
g_j \varphi(x + e) \leq g_j \varphi(x) + |g_j| l|e|.
$$

- $g < 0$: Here, we get

$$
g_j \varphi(x) - g_j l|e| \geq g_j \varphi(x + e) \geq g_j \varphi(x) + g_j l|e|.
$$

  Since $g_j = \text{sign}(g_j)|g_j| = -1 \cdot |g_j|$, the left inequality can be rewritten as

$$
g_j \varphi(x) - g_j l|e| = g_j \varphi(x) + |g_j| l|e| \geq g_j \varphi(x + e)
$$

Hence, the following inequality is valid for all $x \in \mathbb{R}$:

$$
g_j \varphi(x + e) \leq g_j \varphi(x) + |g_j| l|e|
$$

The derivation of

$$
g_j \varphi(x + e) \geq g_j \varphi(x) - |g_j| l|e|
$$

follows the respective other branch of both cases. $\qquad\square$

References

[1] C. McCollough, "Tu-fg-207a-04: Overview of the low dose ct grand challenge," *Medical Physics*, vol. 43, no. 6Part35, pp. 3759–3760, 2016.

[2] I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio, "Maxout networks," *arXiv preprint arXiv:1302.4389*, 2013.

[3] D. L. Parker, "Optimal short scan convolution reconstruction for fan beam ct," *Medical Physics*, vol. 9, no. 2, pp. 254–257, 1982.

[4] J. Staal, M. D. Abràmoff, M. Niemeijer, M. A. Viergever, and B. Van Ginneken, "Ridge-based vessel segmentation in color images of the retina," *IEEE Transactions on Medical Imaging*, vol. 23, no. 4, pp. 501–509, 2004.

[5] K. Zuiderveld, "Graphics gems iv," P. S. Heckbert, Ed., 1994, ch. Contrast Limited Adaptive Histogram Equalization, pp. 474–485.

[6] H. Wu, S. Zheng, J. Zhang, and K. Huang, "Fast End-to-End Trainable Guided Filter," in *CVPR*, 2018, pp. 1838–1847.

[7] Q. Chen, J. Xu, and V. Koltun, "Fast Image Processing With Fully-Convolutional Networks," in *Proc IEEE Int Conf Comput Vis*, 2017, pp. 2497–2506.

[8] A. Odena, V. Dumoulin, and C. Olah, "Deconvolution and checkerboard artifacts," *Distill*, vol. 1, no. 10, p. e3, 2016.

[9] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[10] C. Syben, B. Stimpel, J. Lommen, T. Würfl, A. Dörfler, and A. Maier, "Deriving neural network architectures using precision learning: Parallel-to-fan beam conversion," in *German Conference on Pattern Recognition (GCPR)*, 2018.

[11] L. A. Shepp and B. F. Logan, "The fourier reconstruction of a head section," *IEEE Transactions on nuclear science*, vol. 21, no. 3, pp. 21–43, 1974.

[12] R. Paulavičius and J. Žilinskas, "Analysis of different norms and corresponding lipschitz constants for global optimization," *Technological and Economic Development of Economy*, vol. 12, no. 4, pp. 301–306, 2006.